PREDICTION OF TALEN BINDING SITE FOR GENOME EDITING IN STEM CELLS

A THESIS SUBMITTED IN PARTIAL FULFILMENT

OF THE REQUIREMENT FOR THE DEGREE OF

Bachelor of Technology
in
BIOTECHNOLOGY
By

PRITI AGRAWAL
111BT0492

Under the Supervision of

Dr. MUKESH KUMAR GUPTA

Department of Biotechnology and Medical Engineering

National Institute of Technology Rourkela

Rourkela, Odisha, 769 008, India

May 2015

Certificate

This is to certify that the thesis entitled **—** "**MOLECULAR MODELLING OF GENOME EDITING IN STEM CELLS BY TALENs"** by **PRITI AGRAWAL (111BT0492),** in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in BIOTECHNOLOGY during session 2011-2015 in the Department of Biotechnology and Medical Engineering, National Institute of Technology Rourkela, is an authentic work carried out by her under my supervision and guidance. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/Institute for the award of any degree or diploma.

Place: Rourkela

Date: 11th May 2015

Dr. M. K Gupta
Associate Professor
Biotechnology and Medical Engineering
National Institute of Technology Rourkela-769 008,

ACKNOWLEDGEMENT

On the successful completion of this project, now I have the golden opportunity to express my sincere gratitude to my guide Prof. Mukesh Kumar Gupta (Associate Professor, Department of Biotechnology and Medical Engineering) for his invaluable supervision, guidance and critical motivation during this project. I am thankful to my guide for giving me an opportunity to work on such an interesting and challenging project.

I would like to put forth my sincere gratitude to Mr.Guttula Praveen Kumar and Ms.Tanushree Patra who helped me throughout this project. Without their help I could never be able to understand the aspects of this project.

I also extends my thanks to my family and friends for their constant support and motivation during the time of dismay.

Priti Agrawal

Biotechnology and Medical Engineering

National Institute of Technology Rourkela

Pin-769 008, Odisha (India)

CONTENTS

Page No.

LIST OF TABLES :

LIST OF FIGURES:

ABSTRACT

Transcription activator-like effector (TALE) nucleases (TALENs) have currently risen as a genome editing tool in numerous live forms .Transcription activator-like effector nucleases (TALENs), comprising of an DNA binding domain and a FokI cleavage domain TALENs can behaves as adaptable molecular DNA scissors creating double-strand breaks (DSBs) at desired position of a genome, expanding the effectiveness of genomic modification. The TALE central repeat domains's flexible nature empower to tailor DNA recognition specificity easily and target basically any desired DNA sequence.

In this study we understood the structural basis of interaction between TALEN and DNA and the target binding site of TALEN was predicted. It was found that NH RVD binds Guanine more specifically than NN RVD and accordingly script of TALE NT TOOL 2.0(https://tale-nt.cac.cornell.edu/node/add/talen) was modified. Adh2 and adh1c gene were tested in both the existing and modified script of TALE NT tool 2.0. Experiment shows that a 63 aa CTS in TALEN scaffold ,could lead to effective gene modification in cells of human, when separated by 14-32 bp spacers and accordingly script of TALE NT was changed with the spacer length 15-30.The results of adh2 gene and adh1c gene were verified by online available tools of TALEN LIBRARY RESEARCH (http://www.talenlibrary.net/) and we got affirmative results. Thus, the modified script is expected to increase the efficiency of DNA targeting.

.

CHAPTER ONE

INTRODUCTION

INTRODUCTION

Genome editing, is a type of genetic engineering in which DNA is embedded, supplanted, or expelled from a genome utilizing engineered nucleases. Targeted genome modification (TGM), intervened by engineered nucleases, has beengenerally used to examine gene function and expand biotechnology applications in various organisms.

TALENs, a type of engineered nucleases consist of aTALE DNA binding domain and a FokI cleavage domain.The DNA binding domain, composed of various almost identical tandem repeat arrays, can target any given sequence as per a basic repeat variable di-residue (RVD)-nucleotide recognition code.

As per this code, the HDrepeat specifies C, NG specifies T, NI specifies A, NN specifiesG or A, N* specifies C, IG specifies T, and NS specifiesA, C, G, or T. Albeit numerous common RVDs have been foundfour of them, HD, NN, NI, and NG, represent for75% of the aggregate.

The discriminating venture in TGM is presentation of DNAdouble-strand breaks (DSBs) at desired genomic location. It wasbefore long found that designed nucleases could createDSBs and thus initiate DNA repair to seal the breaksalongside any changes, for example, transformations, insertions,substitutions.

TALE-NT is a unreserved freely available tool for designing pairs of TAL effectors for TAL effector nucleases (TALENs) to target a specific gene sequence.Moreover, TALE-NT gives tools to design single TAL effectors to target to a DNA sequence, and to find binding sites in a DNA sequence for a given TAL effector.

To find out the ideal TALEN building design with efficient cleavage activity and minimal peptide segment, scaffold optimization has been carried out.A methodical study was

done by building 10 diverse TALEN scaffold with different NTSs and CTSs. The DNA cleavage action of every framework was examined against 10 substrates with diverse spacers in a yeast correspondent framework. Taking into account this 10_10network, two TALEN platforms with high DNA cleavage productivity in both yeast and human cells were distinguished.

One bearing a 207 aa NTS and a 31 aa CTS inclines toward target destinations with 10–16 bp spacers while another bearing a 207 aa NTSalso, a 63 aa CTS has most noteworthy proficiency when differentiated by 14–32 bp spacers. Accordingly, the spacer length of the script was modified to 15-30 bp spacer length.

It was additionally observed that NH RVD ties "G" more particularly than NN RVD. Keping this as a main priority the script of TALE NT was adjusted and two genes in particular adh2 and adh1c were run through the current script of TALE NT and through the changed version of TALE NT TOOL. The results were verified using an online tool of TALE NT and TALEN LIBRARY RESOURCE.

Objective:

The objective of this project is as followed

- To understand the structural basis of interaction between Transcription Activator Like Effectors Nucleases and target DNA site..
- To analyze the target prediction and scaffold optimization for genome editing in stem cell by scripting via python programming.

CHAPTER TWO

LITERATURE REVIEW

LITERATURE REVIEW:

TALE STRUCTURE AND RECOGNITION CODE

Structure of TALE includes:

- N terminal region which comprises secretion and translocation signals
- C terminal region consisting of nuclear localization signals (NLS)

an acidic transcription-activation domain (AD)

- A central DNA binding domain (DBD)

DBD consist of amino acid repeat which is 33-35 aa long and containing only 20 amino acids in the last repeat which is thus called as half repeat.

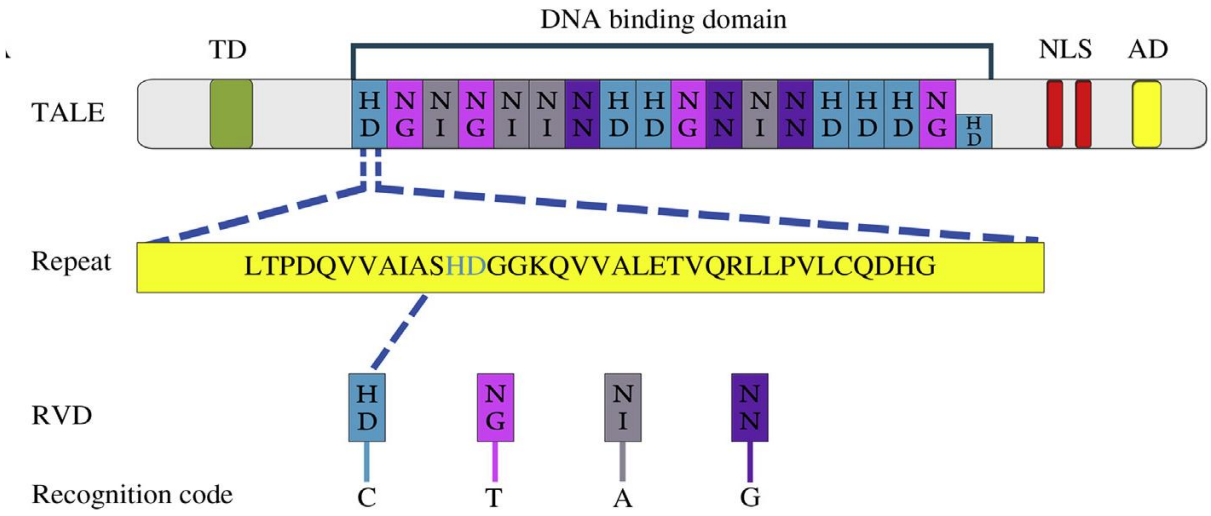


Figure 1. Structure of TALE [1]

The DNA binding specificity of TALE is a characteristics property of repeat variable di-residue (RVD) in the location 12 and 13 of each repeat [2] [13].

RVD	Nucleotide specified by RVD
HD	C
NG	T
NI	A
NN	G or A
N*	C
IG	T
NS	A,C, G or T
NH	G
NK	G

Table 1.RVDs and corresponding specific nucleotide

Albeit numerous RVDs have been found, mainly HD, NN, NI, and NG, represent 75% of the aggregate [6] .NK show poor affinity toward G relative to NH [5]. The RVD sequence determines the specificity of DNA binding. The number of repeat corresponds to the total length of tarset site and RVD determines the nucleotide to be recognized in the target site. Besides all this, a thymine (T) succeeds the recognition sites and this is the basic principle of TALENs for gene targeting. TAL repeat consists of 2 helices joined by a short RVD-containing loop, and all repeats form a right-handed superhelical structure that tracks along the sense strand of the DNA duplex, with the RVDs contacting the major groove [7]. The two hyper-variable residues have distinct biochemical function in the RVD loops.

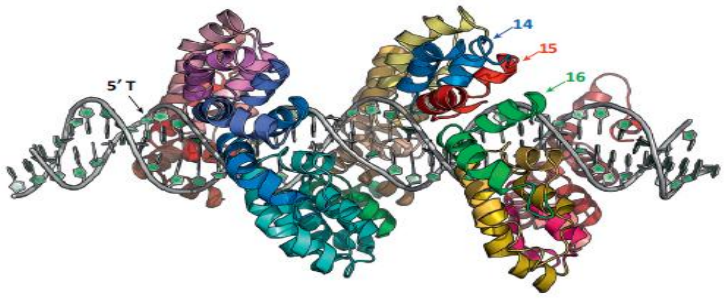


Figure2. TALE-DNA interaction structure (side view)

Figure 3.TALE –DNA interaction structure (top view, along the axis of DNA groove)

The amino acid in position 12 gives stability to the structure by interacting with the residue of repeats through hydrogen bond.The amino acid in position 13 is responsible for the specificity of binding of TALE to specific nucleotide in the target site. [3] [4]

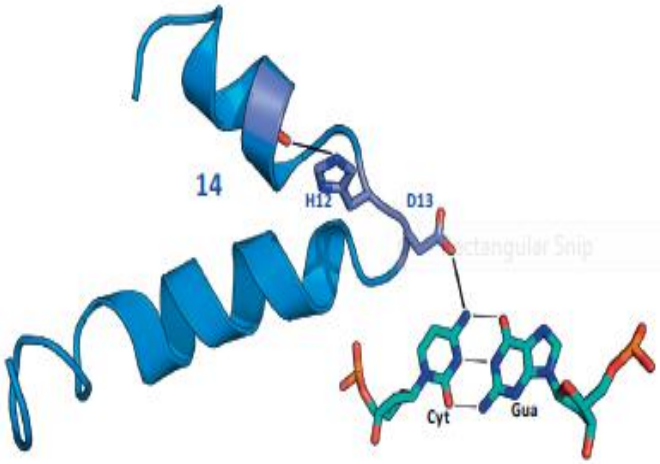


Figure 4.Interaction of RVD 12 and 13 in repeat 14

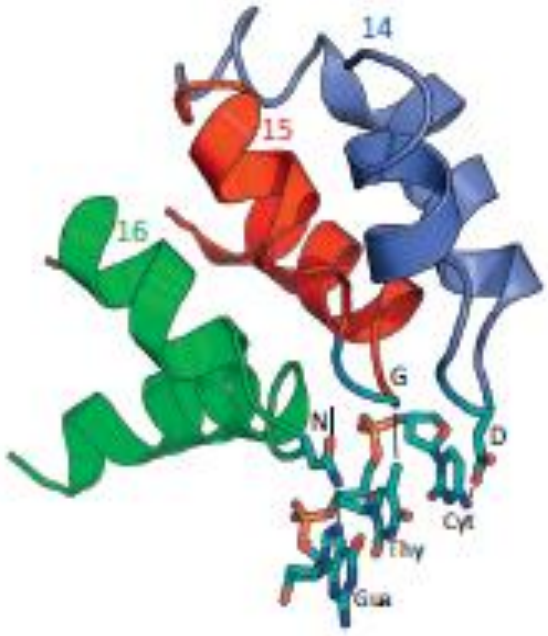


Figure 5. Interaction of DNA with the repeat 14, 15, 16

The residue at position 12 in each RVD arranges far from DNA and interact with the carbonyl at position 8 in the first helix of protein backbone, and hence gives stability to the loop.The residue at position 13 profound into DNA major groove and interacts with the specific nucleotide of DNA sequence [12].

SPECIFICITIES OF RVDs FOR “G”

To check the specificities of RVD, Artificial TALEs (ArtTALs) was constructed where the initial six repeats were from natural TALE Hax3,containing several HD RVDs. There were total of 11.5 RVD where the next 6 were identical RVD. We transfected ArtTAL expression with□-glucuronidase (GUS) reporter constructs was tranfected into leaf cells of Nicotianabenthamiana. This man made setup was used to examine the RVD (known and new).The specificity for G was shown by NH RVD. When compared to NK, the G specific RVD NH was found to be much more efficient [14]. These RVD specificities will allow a highly specific targeting of guanine (NH)and a flexible targeting .

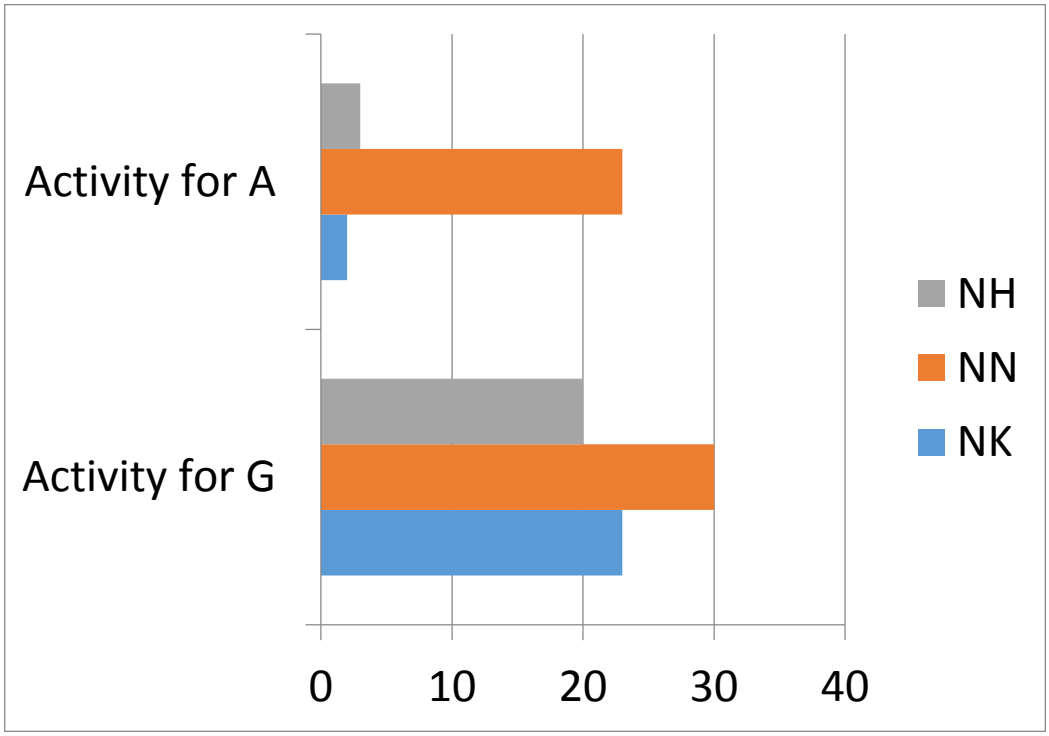


Figure 6: Specificity for G by different RVD

SCAFFOLD OPTIMISATION

TALEs which occurs naturally are transcription activators. NTSs maintains protein secretion signal peptides.Their CTSs comprises of nuclearlocalization signal peptides and a transcription activatordomain [10].From the experiments on yeast assay, between the two TALEN binding sites(NTS 288 aa and CTS 295 aa) the spacer length was determined to be 16–31 bp. Alternatively, the TALENs (287 aa NTS and 231 aa CTS) was efficient with 13–30 bp spacers.

To find out the ideal TALEN building design with most astoundingcleavage proficiency and minimal peptide segment, scaffold optimization has been carried out.A methodical study was done , in which 10 diverse TALEN scaffold with different NTSs and CTSs was built . The DNA cleavage action ofevery framework was examined against 10 substrates with diversespacers in a yeast correspondent framework. Taking into account this 10_10network, two TALEN platforms with high DNA cleavageproductivity in both yeast and human cells were distinguished.Onebearing a 207 aa NTS and a 31 aa CTS inclines toward target destinationswith 10–16 bp spacers while another bearing a 207 aa NTSalso, a 63 aa CTS has most noteworthy proficiency when differentiated by 14–32 bp spacers (Table II) [14][9]

SR NO	NTS	CTS	Spacer(bp)
1	288	295	16-31
2	288	285	16
3	287	231	13-30
4	287	63	15
5	207	63	14-32
6	207	31	10-16
7	153	47	12-21
8	153	17	12
9	136	63	12-20
10	136	28	12-13
11	136	18	13-16

Table 2. Spacer length of different TALENs

CHAPTER THREE

METHODS AND SOFTWARE REQUIREMENTS

METHOD AND SOFTWARE REQUIREMENTS

TALE NT TOOL 2.0

The tools are located at /opt/boglab/talent in the TALE NT software's script file(by default).Following steps were followed to change the location of the tool.

1.In talconfig.py, the value of BASE_DIR was changed to to the full path of the parent folder of boglab_tools

2.In findTAL.py, 'talent' was changed to 'boglab_tools' in the line 'with open(BASE_DIR + "/talent/re_dict_dump", "rb") as re_dict_file".

Boglab tools required Python 2.6+ and Biopython,findRvdTAL doest not work until boglab_tfcount , C libraries are installed, And findPairedRvdTALs.py does not work until cython wrappers from boglab_talesf is installed. "findTAL.py" scripts in boglab_tools corresponds to the TALEN Targeter in TALE NT Tool.

Command to run script:	python findTal.py OPTIONS
Command to find the existing option	Python findTAL.py --help
Command for running findTAL.py :	Python findTAL.py – filter'0' –gspec—min '15' –max '30'—arraymax '20' – arraymin"15"—fasta'filemx4P32.fasta'

Here min represents minimum spacer length, max represents maximum spacer length, arraymin represents minimum RVD sequence length, arraymax represents maximum RVD sequence length, fasta represents path to the file containing input sequences in FASTA format, gspec uses NH instead of NN for G

PYTHON:

Install python by following this command:	Python 3.4 setup.py build Python 3.4 setup.py test Sudo python 3.4 setup.py install

After installation of python, we installed biopthon.

BIOPYTHON:

To install the biopython, zip format of source code was downloaded and source code were unzipped.

For installation, the following commands were typed in the directory	Python setup.py build Python setup.py test Sudo python setup.py install

To test python and biopython installation, the following code were typed in biopytho source codePython setup.py test

Tfcount:

Following softwares were installed to use tfcount:

1.NVIDIA GPU 2.0+

2.CUDA Toolkit 4.2+,

3.libcutils

4.Python 2.6+

5.cython.

To install tfcount, following command were typed in directory	Make Make install Cdcython_wrapper Python setup.py build-ext Python setup.py install

Following commands were typed to run tfcount	python findTALAddOffTargets.py offtargetseq PATH_TO_FASTA_FILE -outpathPATH_TO_TALEN_TARGETER_RES

Libcutils:

Following commands were typed in the directory of libcutils for its installation	Make Make install

FASTA sequence of ADH1C gene:

>GAACTCTGAATGACCCCTGTGGGTTTGAGAGAAGAGAAGCAGGAACTTGAGAGAGGAGGA
AGAGAGAAAGTAATTAAAATGTATCGTTTTAACTTAATATTTAACCGAATGATAGCAAAATC
TTATCTGAAATTGGAAAAGTCAAGGTTTTGAGTGCTGGTTCGGTGCCCATTTCTTTATGATTT
GATAGTCTGAGAAGAATACGACGGGTGTGGCTTAAAAACCTAGATCACGTGTGTAGTTGGA
ATTGGGTGTTATATGAGCAAACAAAATAAATACCTGTGCAACATACCTGCTTTATGCACTCA
AGCAGAGAAGAAATCCACAAGTACTCACCAGCCTCCTGGTCTGCAGAGAAGACAGAATCAA
TATGAGCACAGCAGGAAAAGTAATCAAATGCAAAGCAGCTGTGCTATGGGAGTTAAAGAAA
CCCTTTTCCATTGAGGAGGTAGAGGTTGCACCTCCTAAGGCTCATGAAGTTCGCATTAAGAT
GGTGGCTGCAGGAATCTGTCGTTCAGATGAGCATGTGGTTAGTGGCAACCTGGTGACCCCCC
TTCCTGTGATTTTAGGCCATGAGGCAGCCGGCATCGTGGAAAGTGTTGGAGAAGGGGTGACT
ACAGTCAAACCAGGTGATAAAGTCATCCCGCTCTTTACTCCTCAGTGTGGAAAATGCAGAAT
TTGTAAAAACCCAGAAAGCAACTACTGCTTGAAAAATGATCTAGGCAATCCTCGGGGGACC
CTGCAGGATGGCACCAGGAGGTTCACCTGCAGCGGGAAGCCCATCCACCACTTCGTCGGCGT
CAGCACCTTCTCCCAGTACACAGTGGTGGATGAGAATGCAGTGGCCAAAATTGATGCAGCCT
CGCCCCTGGAGAAAGTCTGCCTCATTGGCTGTGGATTTTCGACTGGTTATGGGTCTGCAGTC
AAAGTTGCCAAGGTCACCCCAGGGTCTACCTGTGCTGTGTTTGGCCTGGGAGGGGTCGGCCT
ATCTGTTGTTATGGGCTGTAAAGCAGCTGGAGCAGCCAGAATCATTGCTGTGGACATCAACA
AGGACAAATTTGCAAAGGCTAAAGAGTTGGGTGCCACTGAATGCATCAACCCTCAAGACTA
CAAGAAACCCATTCAGGAAGTGCTAAAGGAAATGACTGATGGAGGTGTGGATTTTTCGTTTG
AAGTCATCGGTCGGCTTGACACCATGATGGCTTCCCTGTTATGTTGTCATGAGGCATGTGGC
ACAAGTGTCATTGTAGGGGTACCTCCTGATTCCCAGAACCTCTCAATAAACCCTATGCTGCT
ACTGACTGGACGCACGTGGAAAGGAGCTATTTTTGGAGGCTTTAAGAGTAAAGAATCTGTCC
CCAAACTTGTGGCTGACTTTATGGCTAAGAAGTTTTCACTGGATGCATTAATAACAAATATTT
TACCTTTTGAAAAAATAAATGAAGGATTTGACCTGCTTCGTCTGGAAAGAGTATCCGTACCG
TCCTGACGTTTTGAAACAATACAGATGCCTTCCCTTGTAGCAGTTTTCAGCCTCCTCTACCCT
ACATGATCTGGAGCAACAGCTAGGAAATATCATTAATTCTGCTCTTCAGAGATGTTAAAAAT
AAATTACACGTGGGAGCTTTCCAAAGAAATGGAAATTGATGGGAAATTATTTGTCAAGCAA
ATGTTTAAAATCCAAATGAGAACTAAATAAAGTGTTGAACATCAACTGGGGAATTGAAGCC
AATAAACCTTCCTTCTTAACCATTCAAAAAAAAAAAAA

FASTA sequence of ADH2 gene:

>AAATTTAACTGGGTTTCCTTGATTGATTGGTGTTTTAATTAATTGGATGTGTAGGATTGTGG
AGAGTGTTGGCGAAGGTGTGACTGATGTCAAGCCTGGTGACCATGTGCTTCCCATATTCACT
GGGGAGTGCAAGGAGTGCCGGCATTGCAAGTCGGAGGAGAGCAACATGTGTGACCTTCTTA
GGATAAACACTGACAGAGGGGTGATGCTTAATGATGGCAAGTCGAGGTTTTCCATCAATGG
ACAACCCATTTTCCATTTCGTTGGAACTTCAACCTTCAGTGAATACACCGTTCTTCATGTTGG
TTGTCTTGCCAAGATCAATCCAGAGGCTCCACTTGATAAAGTTTGTGTTCTCAGCTGTGGAAT
CTCTACAGGTTAGGATTGGGATTCAAATCCTATCTTATTTATTCCTTTGGCTTGTGTTTCTTAC
TTCTCTCTAATCTAATAAGTAAGAATTGGGTTTATGTGCCTACTGTAGGTCTTGGTGCCACCT
TGAATGTTGCAAAACCCAAGAAGGGTTCAACTGTAGCCATTTTTGGATTAGGAGCTGTGGGG
CTTGCCGTGAGTTCTTCAGCTACCATTCTTTTGTTTATTTCTAGAGCTGCTGCTAATGAGTTAA
ATTGTGTTTTGTTTGACATTATTAGGCCGCGGAAGGGGCTAGAATCGCAGGTGCTTCAAGGA
TTATTGGTGTTGATTTGAACCCTAAGAGATTTGATGAAGGTGAATGCTACTTAAAACATCAA
CCCCAGTTCATTTTTGTTCTCTAGTTATGTAGCTCCTATTCATTGCCTATTTCTTGTAAACAGC
TAGGAAATTTGGTGTGACCGAATTTGTGAACCCAAAGGATTATAACAAGCCTGTTCAAGAGG
TATTTTGATTTTCTACATAAATTGCAATTGCATTGTAGTGGTGAAGATAGAAATATAAAACT
GCAGTTTGGATTAATCTTGCTCTTTCCTCTGTAATTAGGCGATTGCTGAGATGACTGGTGGTG
GAGTAGACCGAAGTGTTGAATGTACAGGAAGTATCATGGCCATGATGTCTGCATTTGAGTGT
GTTCATGATGTAATTTCTATTACAACTTGTTAATGCTGTTATTCCCAAGTATAGGA

Run the fasta sequence of ADH1C and ADH2 gene in the script file of TALE NT tool and in the modified script of TALE NT TOOL.

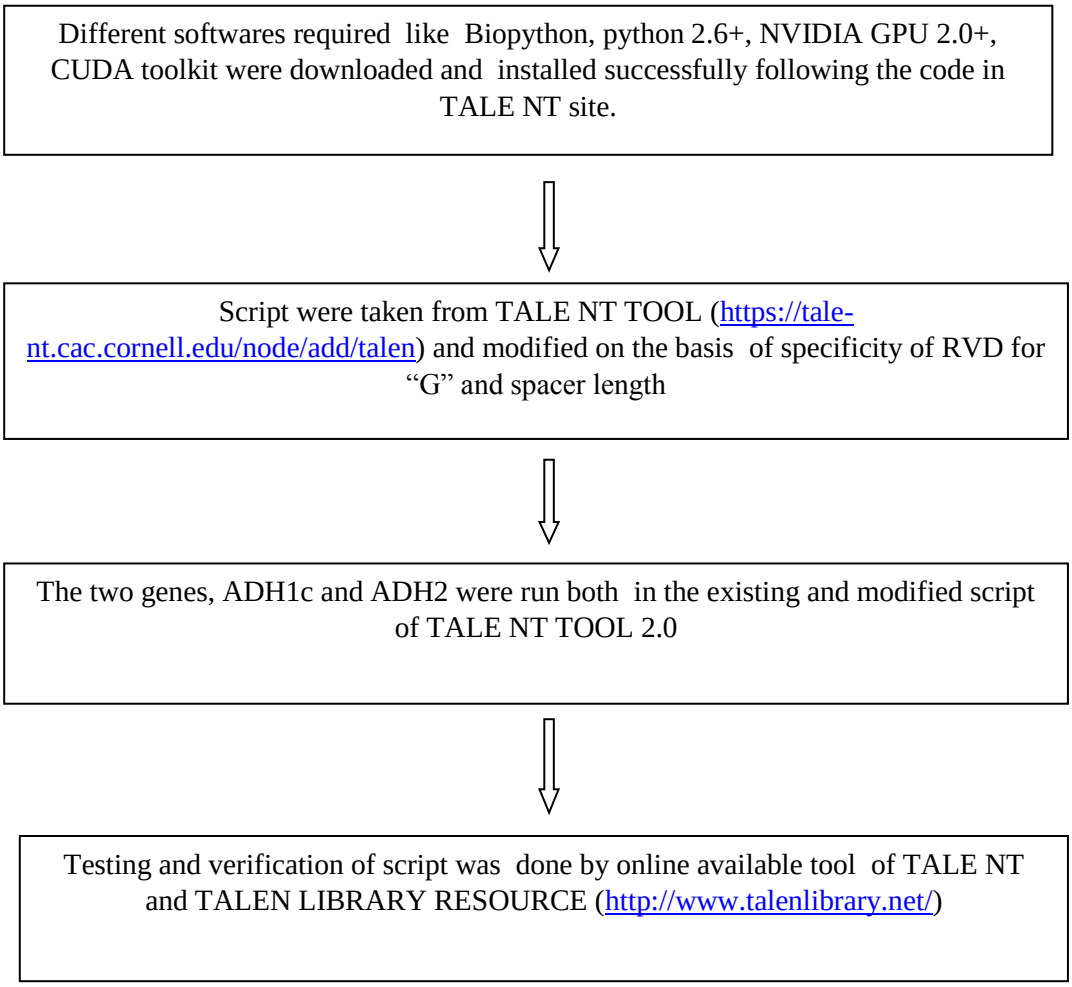
Script of TALEN Targeter was written on the basis of following criteria:

- "T" should preceed before the TALEN recognition site/target site.
- The target site selected should be unique in nature and nucleotide code.
- For each TALEN, there should be at least 14 tandem repeats in the DNA binding site of TALE.
- To stabilize the interaction between TALE –DNA there should be at least four strong

RVDs (HD, NG, NN).

- Furthermore, there should be strong RVD at the termini of the TALE sequence.
- Stretches should be avoided for 3 identical RVDs, as there can be deformation in the structure due to the presence of same consecutive RVD.
- For targeting "G", use NH instead of NN as NH is more specifie toward "G".
- Use the spacer length to be 15-30 bp for scaffold optimisation.
- There should not be any T at position one or position two.
- No "G" should be present at the end of half site of the target sequence.

Flowchart of work done:



CHAPTER THREE

RESULTS AND DISCUSSION

RESULT AND DISCUSSION:

In the Script of TALE NT, the minimum RVD sequence length was 15 and the maximum RVD sequence length was 20. The spacer length was taken to be 15-20 bp. Besides all this, a thymine (T) succeeds the recognition sites, which is the basic principle of TALENs for gene targeting.

In the modified script, the minimum RVD sequence length was 15 and maximum was 20. The spacer length was taken as 15-30 and NH is taken as the RVD which will recognize and interact with "G".For the verification of data from the script, we have tested the genes in the online TALE NT TOOL, and it was found to be same as of the script. This verifies the procedure followed in running the script.

Both the gene ADH1c and ADH2 were tested for both the existing and modified script.

SR NO	TARGET SITE	RVD SEQUENCE
1	T ATATGAGCAAACAAAAT ATGCACTCAAGCAGAG	NI NG NI NG NN NI NN HD NI NINI HD NI NININI NG HD NG HD NG NN HD NG NG NN NI NN NG NN HD NI NG
2	T CCTCTACCCTACATGAT ATTAATTCTGCTCTTCAG	HD HD NG HD NG NI HD HDHD NG NI HD NI NG NN NI NG HD HD NG HD NG NI HD HDHD NG NI HD NI NG NN NI NG
3	T CCTCTACCCTACATGAT ATTAATTCTGCTCTTCAGAG	HD HD NG HD NG NI HD HDHD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG
4	T CCTCTACCCTACATGAT AATTCTGCTCTTCAGAG	HD HD NG HD NG NI HD HDHD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG
5	T CCTCTACCCTACATGATCT ATTAATTCTGCTCTTCAG	HD HD NG HD NG NI HD HDHD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG

Table3. Sequence of RVDs from the existing script of TALE NT for ADH1C gene

SR NO	TARGET SITE	RVD SEQUENCE
1	T CTGAATGACCCCTGT GAAGCAGGAACTTGAGAG	HD NG NH NI NI NG NH NI HD HDHDHD NG NH NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH HD NG NG HD
2	T CTGAATGACCCCTGTG AAGCAGGAACTTGAGAGAGG	HD NG NH NI NI NG NH NI HD HDHDHD NG NH NG NH HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH HD NG NG
3	T CTGAATGACCCCTGTGG AGCAGGAACTTGAGAGAGG	HD NG NH NI NI NG NH NI HD HDHDHD NG NH NG NH NH HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH HD NG
4	T CTGAATGACCCCTGTGG CAGGAACTTGAGAGAGGAGG	HD NG NH NI NI NG NH NI HD HDHDHD NG NH NG NH NH HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH
5	T CTGAATGACCCCTGTGGGT CAGGAACTTGAGAGAGGAGG	HD NG NH NI NI NG NH NI HD HDHDHD NG NH NG NH NHNH NG HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH

Table 4. Sequence of RVDs from the modified script of TALE NT for ADH1C gene

SR NO.	TARGET SITE	RVD SEQUENCE
1	T TTAACTGGGTTTCCT TTAATTAATTGGATGTGT	NG NG NI NI HD NG NH NHNH NG NGNG HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI NG NG NI NI
2	T TTAACTGGGTTTCCTT TAATTAATTGGATGTGTAGG	NG NG NI NI HD NG NH NHNH NG NGNG HD HD NG NG HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI NG NG NI
3	T TTAACTGGGTTTCCTTG AATTAATTGGATGTGTAGG	NG NG NI NI HD NG NH NHNH NG NGNG HD HD NG NG NH HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI NG NG
4	T TTAACTGGGTTTCCTTGAT TTGGATGTGTAGGATTGTGG	NG NG NI NI HD NG NH NHNH NG NGNG HD HD NG NG NH NI NG HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD NI NI
5	T TAACTGGGTTTCCTTGATTG TTGGATGTGTAGGATTGTGG	NG NI NI HD NG NH NHNH NG NGNG HD HD NG NG NH NI NG NG NH HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD NI NI

Table 5. Sequence of RVDs from the modified script of TALEN for ADH2 gene

SR NO.	TARGET SITE	RVD SEQUENCE
1	T CTAGTTATGTAGCTCCT AGGAAATTTGGTGTGACCG	HD NG NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG HD NN NN NG HD NI HD NI HD HD NI NINI NG NGNG HD HD NG
2	T GCAATTGCATTGTAGT ATAAAACTGCAGTTTGG	NN HD NI NI NG NG NN HD NI NG NG NN NG NI NN NG HD HD NI NINI HD NG NN HD NI NN NG NGNGNG NI NG
3	T GCAATTGCATTGTAGT AAAACTGCAGTTTGG	NN HD NI NI NG NG NN HD NI NG NG NN NG NI NN NG HD HD NI NINI HD NG NN HD NI NN NG NGNGNG
4	T AGTTATGTAGCTCCT AGGAAATTTGGTGTGACCG	NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG HD NN NN NG HD NI HD NI HD HD NI NINI NG NGNG HD HD NG
5	T AGTTATGTAGCTCCTAT AGGAAATTTGGTGTGACCG	NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG NI NG HD NN NN NG HD NI HD NI HD HD NI NINI NG NGNG HD HD NG

Table 6. Sequence of RVDs from the existing script of TALE NT for ADH2 gene

CONCLUSION:

The Transcription Activator Like Effector (TALE) structure was understood and the analysis of the interaction between TALE and target DNA site was done followed by modifying the python script of TALE NT TOOL on the basis of "G" specificity parameter and spacer length. Two genes namely ADH2 and ADH1C were tested in the script of TALE NT and results were verified using an online tool of TALE NT TOOL 2.0 (https://tale-nt.cac.cornell.edu/node/add/talen) and TALEN LIBRARY RESOURCE (http://www.talenlibrary.net/). Results from online tool found to be in coordination with the results from the python script.This verifies our TALEN sequence found from the script. And, thus the script with modification parameters will increase the efficiency of specific DNA targeting of TALENs.

REFERENCES:

1.Moscou, M.J., Bogdanove, A.J., 2009. A simple cipher governs DNA recognition by TAL effectors. Science 326, 1501

2.Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S.,Lahaye, T., Nickstadt, A., Bonas, U., 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. Science 326, 15

3. Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.K., Shi, Y., Yan, N.,2012a. Structural basis for sequence-specific recognition of DNA by TALeffectors. Science 335, 720e723.

4. Mak, A.N., Bradley, P., Cernadas, R.A., Bogdanove, A.J., Stoddard, B.L.,2012. The crystal structure of TAL effector PthXo1 bound to its DNAtarget. Science 335, 716e719.

5.Christian, M.L. et al. (2012) Targeting G with TAL effectors: acomparison of activities of TALENs constructed with NN and NKrepeat variable di-residues. PLoS ONE 7, e45383

6. Streubel, J. et al. (2012) TAL effector RVD specificities and efficiencies. Nat. Biotechnol. 30, 593–595

7. Murakami, M.T. et al. (2010) The repeat domain of the type IIIeffector protein PthA shows a TPR-like structure and undergoesconformational changes upon DNA interaction. Proteins 78, 3386–3395

8.Christian ML, Demorest ZL, Starker CG, Osborn MJ, Nyquist MD, Zhang, Carlson DF, Bradley P, Bogdanove AJ, Voytas DF. 2012. Targeting Gwith TAL effectors: A comparison of activities of TALENs constructedwith NN and NK repeat variable di-residues. PLoS ONE 7:e45383

9. White FF, Potnis N, Jones JB, Koebnik R. 2009. The type III effectors of Xanthomonas. Mol Plant Pathol 10:749–7766.

10.Ning Sun,1Huimin Zhao (2013) Transcription Activator-Like Effector Nucleases (TALENs): A Highly Efficient and Versatile Tool for Genome Editing 2-4

11. Chuanxian Wei a,b, Jiyong Liu a, Zhongsheng Yu a,b, Bo Zhang c,*, Guanjun Gao d,*, Renjie Jiao a2013 TALEN or Cas9 e Rapid, Efficient and Specific Choices for Genome Modifications 3-6

12.Kunling Chen, CaixiaGaoTALENs: Customizable Molecular DNA Scissors for GenomeEngineering of Plants 2-5

13.Sun N, Liang J, Abil Z, Zhao H. 2012b. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. MolBiosyst 8:1255–1263.

14.Streubel, J., Blucher, C., Landgraf, A., Boch, J., 2012. TAL effector RVDspecificities and efficiencies. Nat. Biotechnol. 30, 593e595.

15. Valton, J., Dupuy, A., Daboussi, F., Thomas, S., Marechal, A., Macmaster, R., Melliand, K., Juillerat, A., Duchateau, P., 2012. Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. J. Biol. Chem. 287, 38427e38432.

16. Tesson L, Usal C, Menoret S, Leung E, Niles BJ, Remy S, Santiago Y, Vincent AI, Meng X, Zhang L, Gregory PD, Anegon I, Cost GJ. 2011. Knockout rats generated by embryo microinjection of TALENs. Nat Biotechnol 29:695–696.

17. Wood AJ, Lo TW, Zeitler B, Pickle CS, Ralston EJ, Lee AH, Amora R, Miller JC, Leung E, Meng X, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Meyer BJ. 2011. Targeted genome editing across species using ZFNs and TALENs. Science 333:307.

18.Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I,Beausejour CM, Waite AJ, Wang NS, Kim KA, Gregory PD, Pabo CO, Rebar EJ. 2007. An improved zinc finger nuclease architecture for highly specific genome editing. Nat Biotechnol 25:778–785.

19.Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE,Leung E, Hinkley SJ, Dulay GP, Hua KL, Ankoudinova I, Cost GJ,Urnov FD, Zhang HS, Holmes MC, Zhang L, Gregory PD, Rebar EJ.2011. A TALE nuclease architecture for efficient genome editing. Nat Biotechnol 29:143–148.

20.Moore FE, Reyon D, Sander JD, Martinez SA, Blackburn JS, Khayter C, Ramirez CL, Joung JK, Langenau DM. 2012. Improved somatic mutagenesis in zebrafish using transcription activator-like effector nucleases (TALENs). PLoS ONE 7:e37877.

21.Morbitzer R, Romer P, Boch J, Lahaye T. 2010. Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. Proc Natl Acad Sci USA 107:21617–21622.

22.Morbitzer R, Elsaesser J, Hausner J, Lahaye T. 2011. Assembly of custom TALE-type DNA binding domains by modular cloning. Nucleic Acids Res 39:5790–5799

ANNEXURE:

Adh2 gene (Existing script)

table_ignores:TAL1 length,TAL2 length,Spacer length

options_used:array_min = 15, array_max = 20, spacer_min = 15, spacer_max = 30, upstream_base = T, No T at position 1, No A at position 1, Sites must end in a T, Sites may not end in G/NN, Base composition rules enforced

Sequence Name TAL1 start TAL2 start TAL1 length TAL2 length Spacer length Spacer range TAL1 RVDs TAL2 RVDs Plus strand sequence Unique_RE_sites_in_spacer

gi|156074003|gb|EF520859.1| 769 832 17 19 28 786-813 HD NG NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG HD HD NG T CTAGTTATGTAGCTCCT attcattgcctatttcttgtaaacagct AGGAAATTTGGTGTGACCG A BsrDI:GCAATG|CATTGC

gi|156074003|gb|EF520859.1| 898 945 16 17 15 914-928 NN HD NI NI NG NG NN HD NI NG NG NN NG NI NN NG HD HD NI NI NI HD NG NN HD NI NN NG NG NG NG NI NG T GCAATTGCATTGTAGT ggtgaagatagaaat ATAAAACTGCAGTTTGG A MboII:GAAGA|TCTTC

gi|156074003|gb|EF520859.1| 898 945 16 15 17 914-930 NN HD NI NI NG NG NN HD NI NG NG NN NG NI NN NG HD HD NI NI NI HD NG NN HD NI NN NG NG NG NG T GCAATTGCATTGTAGT ggtgaagatagaaatat AAAACTGCAGTTTGG A MboII:GAAGA|TCTTC

gi|156074003|gb|EF520859.1| 771 832 15 19 28 786-813 NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG HD HD NG T AGTTATGTAGCTCCT attcattgcctatttcttgtaaacagct AGGAAATTTGGTGTGACCG A BsrDI:GCAATG|CATTGC

gi|156074003|gb|EF520859.1| 771 832 17 19 26 788-813 NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG HD HD NG T AGTTATGTAGCTCCTAT tcattgcctatttcttgtaaacagct AGGAAATTTGGTGTGACCG A BsrDI:GCAATG|CATTGC

gi|156074003|gb|EF520859.1| 771 832 17 16 29 788-816 NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG T AGTTATGTAGCTCCTAT tcattgcctatttcttgtaaacagctagg AAATTTGGTGTGACCG A BsrDI:GCAATG|CATTGC

gi|156074003|gb|EF520859.1| 771 832 17 15 30 788-817 NI NN NG NG NI NG NN NG NI NN HD NG HD HD NG NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG T AGTTATGTAGCTCCTAT tcattgcctatttcttgtaaacagctagga AATTTGGTGTGACCG A BsrDI:GCAATG|CATTGC

gi|156074003|gb|EF520859.1| 775 832 17 19 22 792-813 NI NG NN NG NI NN HD NG HD HD NG NI NG NG HD NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG HD HD NG T ATGTAGCTCCTATTCAT tgcctatttcttgtaaacagct AGGAAATTTGGTGTGACCG A none

gi|156074003|gb|EF520859.1| 775 832 17 16 25 792-816 NI NG NN NG NI NN HD NG HD HD NG NI NG NG HD NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG T ATGTAGCTCCTATTCAT tgcctatttcttgtaaacagctagg AAATTTGGTGTGACCG A none

gi|156074003|gb|EF520859.1| 775 832 17 15 26 792-817 NI NG NN NG NI NN HD NG HD HD NG NI NG NG HD NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG T ATGTAGCTCCTATTCAT tgcctatttcttgtaaacagctagga AATTTGGTGTGACCG A none

gi|156074003|gb|EF520859.1| 264 318 17 17 21 281-301 HD NN NG NG NN NN NI NI HD NG NG HD NI NI HD HD NG NN NN HD NI NI NN NI HD NI NI HD HD NI NI HD NI NG T CGTTGGAACTTCAACCT tcagtgaatacaccgttcttc ATGTTGGTTGTCTTGCC A MboII:GAAGA|TCTTC

gi|156074003|gb|EF520859.1| 264 321 17 20 21 281-301 HD NN NG NG NN NN NI NI HD NG NG HD NI NI HD HD NG HD NG NG NN NN HD NI NI NN NI HD NI NI HD HD NI NI HD NI NG T CGTTGGAACTTCAACCT tcagtgaatacaccgttcttc ATGTTGGTTGTCTTGCCAAG A MboII:GAAGA|TCTTC

gi|156074003|gb|EF520859.1| 264 318 18 17 20 282-301 HD NN NG NG NN NN NI NI HD NG NG HD NI NI HD HD NG NG NN NN HD NI NI NN NI HD NI NI HD HD NI NI HD NI NG T CGTTGGAACTTCAACCTT cagtgaatacaccgttcttc ATGTTGGTTGTCTTGCC A MboII:GAAGA|TCTTC

gi|156074003|gb|EF520859.1| 264 321 18 20 20 282-301 HD NN NG NG NN NN NI NI HD NG NG HD NI NI HD HD NG NG HD NG NG NN NN HD NI NI NN NI HD NI NI HD HD NI NI HD NI NG T CGTTGGAACTTCAACCTT cagtgaatacaccgttcttc ATGTTGGTTGTCTTGCCAAG A MboII:GAAGA|TCTTC

gi|156074003|gb|EF520859.1| 777 832 15 19 22 792-813 NN NG NI NN HD NG HD HD NG NI NG NG HD NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG HD HD NG T GTAGCTCCTATTCAT tgcctatttcttgtaaacagct AGGAAATTTGGTGTGACCG A none

gi|156074003|gb|EF520859.1| 777 832 15 16 25 792-816 NN NG NI NN HD NG HD HD NG NI NG NG HD NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG T GTAGCTCCTATTCAT tgcctatttcttgtaaacagctagg AAATTTGGTGTGACCG A none

gi|156074003|gb|EF520859.1| 777 832 15 15 26 792-817 NN NG NI NN HD NG HD HD NG NI NG NG HD NI NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG T GTAGCTCCTATTCAT tgcctatttcttgtaaacagctagga AATTTGGTGTGACCG A none

gi|156074003|gb|EF520859.1| 777 832 20 19 17 797-813 NN NG NI NN HD NG HD HD NG NI NG NG HD NI NG NG NN HD HD NG HD NN NN NG HD NI HD NI HD HD NI NI NI NG NG NG HD HD NG T GTAGCTCCTATTCATTGCCT atttcttgtaaacagct AGGAAATTTGGTGTGACCG A none

ADH2 GENE (modified script)

table_ignores:TAL1 length,TAL2 length,Spacer length

options_used:array_min = 15, array_max = 20, spacer_min = 15, spacer_max = 30, upstream_base = T

Sequence Name Cut Site TAL1 start TAL2 start TAL1 length TAL2 length Spacer length Spacer range TAL1 RVDs TAL2 RVDs Plus strand sequence Unique RE sites in spacer % RVDs HD or NN/NH

gi|156074003|gb|EF520859.1| 26 4 51 15 18 15 19-33 NG NG NI NI HD NG NH NH NH NG NG NG HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI NG NG NI NI T TTAACTGGGTTTCCT tgattgattggtgtt TTAATTAATTGGATGTGT A none 30

gi|156074003|gb|EF520859.1| 27 4 54 16 20 15 20-34 NG NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI NG NG NI T TTAACTGGGTTTCCTT gattgattggtgttt TAATTAATTGGATGTGTAGG A none 33

gi|156074003|gb|EF520859.1| 28 4 54 17 19 15 21-35 NG NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI NG NG T TTAACTGGGTTTCCTTG attgattggtgtttt AATTAATTGGATGTGTAGG A none 36

gi|156074003|gb|EF520859.1| 29 4 54 18 18 15 22-36 NG NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI NG T TTAACTGGGTTTCCTTGA ttgattggtgtttta ATTAATTGGATGTGTAGG A none 36

gi|156074003|gb|EF520859.1| 30 4 54 19 17 15 23-37 NG NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI NG HD HD NG NI HD NI HD NI NG HD HD NI NI NG NG NI NI T TTAACTGGGTTTCCTTGAT tgattggtgttttaa TTAATTGGATGTGTAGG A none 36

gi|156074003|gb|EF520859.1| 31 4 61 17 20 21 21-41 NG NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD NI NI T TTAACTGGGTTTCCTTG attgattggtgttttaattaa TTGGATGTGTAGGATTGTGG A PacI:TTAATTAA 43

gi|156074003|gb|EF520859.1| 32 4 61 19 20 19 23-41 NG NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI NG HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD NI NI T TTAACTGGGTTTCCTTGAT tgattggtgttttaattaa TTGGATGTGTAGGATTGTGG A PacI:TTAATTAA 41

gi|156074003|gb|EF520859.1| 33 5 61 20 20 17 25-41 NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD NI NI T TAACTGGGTTTCCTTGATTG attggtgttttaattaa TTGGATGTGTAGGATTGTGG A PacI:TTAATTAA 43

gi|156074003|gb|EF520859.1| 34 6 61 20 20 16 26-41 NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD NI NI T AACTGGGTTTCCTTGATTGA ttggtgttttaattaa TTGGATGTGTAGGATTGTGG A PacI:TTAATTAA 43

gi|156074003|gb|EF520859.1| 35 6 63 20 20 18 26-43 NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI HD NG HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD T AACTGGGTTTCCTTGATTGA ttggtgttttaattaatt GGATGTGTAGGATTGTGGAG A AseI:ATTAAT PacI:TTAATTAA 45

gi|156074003|gb|EF520859.1| 36 10 63 19 20 15 29-43 NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI NG NG NH HD NG HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD HD T GGGTTTCCTTGATTGATTG gtgttttaattaatt GGATGTGTAGGATTGTGGAG A AseI:ATTAAT PacI:TTAATTAA 46

gi|156074003|gb|EF520859.1| 37 10 63 20 19 15 30-44 NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI NG NG NH NH HD NG HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG HD T GGGTTTCCTTGATTGATTGG tgttttaattaattg GATGTGTAGGATTGTGGAG A AseI:ATTAAT PacI:TTAATTAA 46

gi|156074003|gb|EF520859.1| 38 10 63 20 18 16 30-45 NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI NG NG NH NH HD NG HD HD NI HD NI NI NG HD HD NG NI HD NI HD NI NG T GGGTTTCCTTGATTGATTGG tgttttaattaattgg ATGTGTAGGATTGTGGAG A AseI:ATTAAT PacI:TTAATTAA 45

gi|156074003|gb|EF520859.1| 39 5 73 20 20 29 25-53 NG NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH HD NH HD HD NI NI HD NI HD NG HD NG HD HD NI HD NI NI NG HD T TAACTGGGTTTCCTTGATTG attggtgttttaattaattggatgtgtag GATTGTGGAGAGTGTTGGCG A AseI:ATTAAT FokI:GGATG|CATCC BtsCI:GGATG|CATCC PacI:TTAATTAA 48

gi|156074003|gb|EF520859.1| 40 6 73 20 20 28 26-53 NI NI HD NG NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI HD NH HD HD NI NI HD NI HD NG HD NG HD HD NI HD NI NI NG HD T AACTGGGTTTCCTTGATTGA ttggtgttttaattaattggatgtgtag GATTGTGGAGAGTGTTGGCG A AseI:ATTAAT FokI:GGATG|CATCC BtsCI:GGATG|CATCC PacI:TTAATTAA 48

gi|156074003|gb|EF520859.1| 41 10 73 19 20 25 29-53 NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI NG NG NH HD NH HD HD NI NI HD NI HD NG HD NG HD HD NI HD NI NI NG HD T GGGTTTCCTTGATTGATTG gtgttttaattaattggatgtgtag GATTGTGGAGAGTGTTGGCG A AseI:ATTAAT FokI:GGATG|CATCC BtsCI:GGATG|CATCC PacI:TTAATTAA 49

gi|156074003|gb|EF520859.1| 42 10 73 20 20 24 30-53 NH NH NH NG NG NG HD HD NG NG NH NI NG NG NH NI NG NG NH NH HD NH HD HD NI NI HD NI HD NG HD NG HD HD NI HD NI NI NG HD T GGGTTTCCTTGATTGATTGG tgttttaattaattggatgtgtag GATTGTGGAGAGTGTTGGCG A AseI:ATTAAT FokI:GGATG|CATCC BtsCI:GGATG|CATCC PacI:TTAATTAA 50

ADH1C GENE (Existing script)

table_ignores:TAL1 length,TAL2 length,Spacer length

options_used:array_min = 15, array_max = 20, spacer_min = 15, spacer_max = 30, upstream_base = T, No T at position 1, No A at position 1, Sites must end in a T, Sites may not end in G/NN, Base composition rules enforced

Sequence Name TAL1 start TAL2 start TAL1 length TAL2 length Spacer length Spacer range TAL1 RVDs TAL2 RVDs Plus strand sequence Unique_RE_sites_in_spacer

gi|589269186|ref|NM_000669.4| 256 314 17 16 26 273-298 NI NG NI NG NN NI NN HD NI NI NI HD NI NI NI NI NG HD NG HD NG NN HD NG NG NN NI NN NG NN HD NI NG T ATATGAGCAAACAAAAT aaatacctgtgcaacatacctgcttt ATGCACTCAAGCAGAG A BfuAI:ACCTGC|GCAGGT BspMI:ACCTGC|GCAGGT

gi|589269186|ref|NM_000669.4| 1537 1595 17 18 24 1554-1577 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG T CCTCTACCCTACATGAT ctggagcaacagctaggaaatatc ATTAATTCTGCTCTTCAG A BfaI:CTAG BpmI:CTGGAG|CTCCAG

gi|589269186|ref|NM_000669.4| 1537 1597 17 20 24 1554-1577 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG T CCTCTACCCTACATGAT ctggagcaacagctaggaaatatc ATTAATTCTGCTCTTCAGAG A BfaI:CTAG BpmI:CTGGAG|CTCCAG

gi|589269186|ref|NM_000669.4| 1537 1597 17 17 27 1554-1580 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG T CCTCTACCCTACATGAT ctggagcaacagctaggaaatatcatt AATTCTGCTCTTCAGAG A BfaI:CTAG BpmI:CTGGAG|CTCCAG

gi|589269186|ref|NM_000669.4| 1537 1597 17 16 28 1554-1581 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG T CCTCTACCCTACATGAT ctggagcaacagctaggaaatatcatta ATTCTGCTCTTCAGAG A BfaI:CTAG BpmI:CTGGAG|CTCCAG

gi|589269186|ref|NM_000669.4| 1537 1595 19 18 22 1556-1577 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG T CCTCTACCCTACATGATCT ggagcaacagctaggaaatatc ATTAATTCTGCTCTTCAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 1537 1597 19 20 22 1556-1577 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG T CCTCTACCCTACATGATCT ggagcaacagctaggaaatatc ATTAATTCTGCTCTTCAGAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 1537 1597 19 17 25 1556-1580 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG T CCTCTACCCTACATGATCT ggagcaacagctaggaaatatcatt AATTCTGCTCTTCAGAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 1537 1597 19 16 26 1556-1581 HD HD NG HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG T CCTCTACCCTACATGATCT ggagcaacagctaggaaatatcatta ATTCTGCTCTTCAGAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 258 314 15 16 26 273-298 NI NG NN NI NN HD NI NI NI HD NI NI NI NI NG HD NG HD NG NN HD NG NG NN NI NN NG NN HD NI NG T ATGAGCAAACAAAAT aaatacctgtgcaacatacctgcttt ATGCACTCAAGCAGAG A BfuAI:ACCTGC|GCAGGT BspMI:ACCTGC|GCAGGT

gi|589269186|ref|NM_000669.4| 1131 1191 16 20 25 1147-1171 NN HD NG NI NI NI NN NN NI NI NI NG NN NI HD NG NN NG HD NI NI NN HD HD NN NI HD HD NN NI NG NN NI HD NG NG T GCTAAAGGAAATGACT gatggaggtgtggatttttcgtttg AAGTCATCGGTCGGCTTGAC A none

gi|589269186|ref|NM_000669.4| 1540 1595 16 18 22 1556-1577 HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG T CTACCCTACATGATCT ggagcaacagctaggaaatatc ATTAATTCTGCTCTTCAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 1540 1597 16 20 22 1556-1577 HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG NI NI NG T CTACCCTACATGATCT ggagcaacagctaggaaatatc ATTAATTCTGCTCTTCAGAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 1540 1597 16 17 25 1556-1580 HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG NG T CTACCCTACATGATCT ggagcaacagctaggaaatatcatt AATTCTGCTCTTCAGAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 1540 1597 16 16 26 1556-1581 HD NG NI HD HD HD NG NI HD NI NG NN NI NG HD NG HD NG HD NG NN NI NI NN NI NN HD NI NN NI NI NG T CTACCCTACATGATCT ggagcaacagctaggaaatatcatta ATTCTGCTCTTCAGAG A BfaI:CTAG

gi|589269186|ref|NM_000669.4| 5 50 15 15 16 20-35 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD NG HD NI NI NN NG NG HD HD NG NN HD NG NG T CTGAATGACCCCTGT gggtttgagagaagag AAGCAGGAACTTGAG A none

gi|589269186|ref|NM_000669.4| 5 52 15 17 16 20-35 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD NG HD NG HD NI NI NN NG NG HD HD NG NN HD NG NG T CTGAATGACCCCTGT gggtttgagagaagag AAGCAGGAACTTGAGAG A none

gi|589269186|ref|NM_000669.4| 5 55 15 20 16 20-35 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD HD NG HD NG HD NG HD NI NI NN NG NG HD HD NG NN HD NG NGT CTGAATGACCCCTGT gggtttgagagaagag AAGCAGGAACTTGAGAGAGG A none

gi|589269186|ref|NM_000669.4| 5 52 15 16 17 20-36 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD NG HD NG HD NI NI NN NG NG HD HD NG NN HD NG T CTGAATGACCCCTGT gggtttgagagaagaga AGCAGGAACTTGAGAG Anone

gi|589269186|ref|NM_000669.4| 5 55 15 19 17 20-36 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD HD NG HD NG HD NG HD NI NI NN NG NG HD HD NG NN HD NG T CTGAATGACCCCTGT gggtttgagagaagaga AGCAGGAACTTGAGAGAGG A none

gi|589269186|ref|NM_000669.4| 5 55 15 16 20 20-39 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD HD NG HD NG HD NG HD NI NI NN NG NG HD HD NG T CTGAATGACCCCTGT gggtttgagagaagagaagc AGGAACTTGAGAGAGG A none

gi|589269186|ref|NM_000669.4| 5 58 15 19 20 20-39 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD HD NG HD HD NG HD NG HD NG HD NI NI NN NG NG HD HD NG T CTGAATGACCCCTGT gggtttgagagaagagaagc AGGAACTTGAGAGAGGAGG A none

gi|589269186|ref|NM_000669.4| 5 58 15 16 23 20-42 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD HD NG HD HD NG HD NG HD NG HD NI NI NN NG NG T CTGAATGACCCCTGT gggtttgagagaagagaagcagg AACTTGAGAGAGGAGG A none

gi|589269186|ref|NM_000669.4| 5 58 15 15 24 20-43 HD NG NN NI NI NG NN NI HD HD HD HD NG NN NG HD HD NG HD HD NG HD NG HD NG HD NI NI NN NG T CTGAATGACCCCTGT gggtttgagagaagagaagcagga ACTTGAGAGAGGAGG A none

ADH1C GENE (Modified Script)

table_ignores:TAL1 length,TAL2 length,Spacer length

options_used:array_min = 15, array_max = 20, spacer_min = 15, spacer_max = 30, upstream_base = T

Sequence Name Cut Site TAL1 start TAL2 start TAL1 length TAL2 length Spacer length Spacer range TAL1 RVDs TAL2 RVDs Plus strand sequence Unique RE sites in spacer % RVDs HD or NN/NH

gi|589269186|ref|NM_000669.4| 27 5 52 15 18 15 20-34 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH HD NG NG HD T CTGAATGACCCCTGT gggtttgagagaaga GAAGCAGGAACTTGAGAG A none 52

gi|589269186|ref|NM_000669.4| 28 5 55 16 20 15 21-35 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG NH HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH HD NG NG T CTGAATGACCCCTGTG ggtttgagagaagag AAGCAGGAACTTGAGAGAGG A none 53

gi|589269186|ref|NM_000669.4| 29 5 55 17 19 15 22-36 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH HD NG T CTGAATGACCCCTGTGG gtttgagagaagaga AGCAGGAACTTGAGAGAGG A none 56

gi|589269186|ref|NM_000669.4| 30 5 58 17 20 17 22-38 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH T CTGAATGACCCCTGTGG gtttgagagaagagaag CAGGAACTTGAGAGAGGAGG A none 56

gi|589269186|ref|NM_000669.4| 31 5 58 19 20 15 24-38 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG NH T CTGAATGACCCCTGTGGGT ttgagagaagagaag CAGGAACTTGAGAGAGGAGG A none 56

gi|589269186|ref|NM_000669.4| 32 5 59 20 20 15 25-39 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG NG HD HD NG T CTGAATGACCCCTGTGGGTT tgagagaagagaagc AGGAACTTGAGAGAGGAGGA A none 53

gi|589269186|ref|NM_000669.4| 33 5 61 20 20 17 25-41 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG HD NG NG HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG NG HD T CTGAATGACCCCTGTGGGTT tgagagaagagaagcag GAACTTGAGAGAGGAGGAAG A none 53

gi|589269186|ref|NM_000669.4| 34 7 61 20 20 15 27-41 NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH HD NG NG HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG NG HD T GAATGACCCCTGTGGGTTTG agagaagagaagcag GAACTTGAGAGAGGAGGAAG A none 53

gi|589269186|ref|NM_000669.4| 35 7 63 20 20 17 27-43 NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH HD NG HD NG NG HD HD NG HD HD NG HD NG HD NG HD NI NI NH NG T GAATGACCCCTGTGGGTTTG agagaagagaagcagga ACTTGAGAGAGGAGGAAGAG A none 53

gi|589269186|ref|NM_000669.4| 36 7 65 20 20 19 27-45 NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH HD NG HD NG HD NG NG HD HD NG HD HD NG HD NG HD NG HD NI NI T GAATGACCCCTGTGGGTTTG agagaagagaagcaggaac TTGAGAGAGGAGGAAGAGAG A none 53

gi|589269186|ref|NM_000669.4| 37 7 66 20 20 20 27-46 NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH NG HD NG HD NG HD NG NG HD HD NG HD HD NG HD NG HD NG HD NI T GAATGACCCCTGTGGGTTTG agagaagagaagcaggaact TGAGAGAGGAGGAAGAGAGA A none 53

gi|589269186|ref|NM_000669.4| 38 11 65 20 20 15 31-45 NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH NI NH NI NH HD NG HD NG HD NG NG HD HD NG HD HD NG HD NG HD NG HD NI NI T GACCCCTGTGGGTTTGAGAG aagagaagcaggaac TTGAGAGAGGAGGAAGAGAG A none 55

gi|589269186|ref|NM_000669.4| 39 11 66 20 20 16 31-46 NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH NI NH NI NH NG HD NG HD NG HD NG NG HD HD NG HD HD NG HD NG HD NG HD NI T GACCCCTGTGGGTTTGAGAG aagagaagcaggaact TGAGAGAGGAGGAAGAGAGA A none 55

gi|589269186|ref|NM_000669.4| 40 5 74 20 20 30 25-54 HD NG NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG NI NI NG NG NI HD NG NG NG HD NG HD NG HD NG NG HD HD NG HD T CTGAATGACCCCTGTGGGTT tgagagaagagaagcaggaacttgagagag GAGGAAGAGAGAAAGTAATT A BpuEI:CTTGAG|CTCAAG SmlI:CTYRAG 45

gi|589269186|ref|NM_000669.4| 41 11 70 20 20 20 31-50 NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH NI NH NI NH NI HD NG NG NG HD NG HD NG HD NG NG HD HD NG HD HD NG HD NG T GACCCCTGTGGGTTTGAGAG aagagaagcaggaacttgag AGAGGAGGAAGAGAGAAAGT A BpuEI:CTTGAG|CTCAAG SmlI:CTYRAG 53

gi|589269186|ref|NM_000669.4| 42 7 76 20 20 30 27-56 NH NI NI NG NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH NG NG NI NI NG NG NI HD NG NG NG HD NG HD NG HD NG NG HD HD T GAATGACCCCTGTGGGTTTG agagaagagaagcaggaacttgagagagga GGAAGAGAGAAAGTAATTAA A BpuEI:CTTGAG|CTCAAG MnlI:CCTC|GAGG SmlI:CTYRAG 43

gi|589269186|ref|NM_000669.4| 43 11 74 20 20 24 31-54 NH NI HD HD HD HD NG NH NG NH NH NH NG NG NG NH NI NH NI NH NI NI NG NG NI HD NG NG NG HD NG HD NG HD NG NG HD HD NG HD T GACCCCTGTGGGTTTGAGAG aagagaagcaggaacttgagagag GAGGAAGAGAGAAAGTAATT A BpuEI:CTTGAG|CTCAAG SmlI:CTYRAG 48